文献与数据学报
第6卷 第3期 2024年9月

建设新时代中国特色中文人工智能语料库的思考

李 栋
（中国社会科学院图书馆，北京 100732）

摘 要：[目的/意义]为争取国际舆论主动权、引导权和话语权，必须高度重视我国人工智能语料库建设。[方法/过程]通过分析生成式人工智能交互内容在舆论传播、文化认同、意识形态等方面的潜在风险，研究当前人工智能语料库发展的瓶颈与困难，在此基础上提出关于中文人工智能语料库建设的建议。[结果/结论]将建设新时代中国特色中文人工智能语料库作为国家战略统筹推进，尽快突破瓶颈，从完善语料库政策制度设计、坚持语料内容正确的政治方向、坚持语料库的公益开放特色、建设面向世界的语料库四个方面着手建设新时代中国特色中文人工智能语料库，为人工智能时代的到来奠定数据基础。

关键词：人工智能 语料库 公益开放
分类号：G250.74 TP391.1 TP18
DOI：10.31193/SSAP.J.ISSN.2096-6695.2024.03.03

0 引 言

随着以ChatGPT为代表的生成式人工智能（Artificial Intelligence Generated Content，简称AIGC）产品在2022年底推出，这种以大模型、大数据和大算力为支撑，拥有接近人类水平的语言理解和生成能力的智能技术，改变了人们学习、工作和生活的固有模式，在全球范围产生巨大影响[1]。此类大语言模型通过连接海量语料库来训练自身模型，以实现更好的自然语言理解。模型训练是否成熟直接取决于用到的语料库是否全面，交互结果的政治立场和倾向主要取决于训练语料库内容。随着生成式人工智能的广泛使用，人们发现在交互内容方面存在不少风险和问题，主要表现在，大模型训练中使用西方主流舆论观点、互联网信息、外文语料库时所带来的民族偏见和种族歧视信息[2]、话语导向和意识形态偏见[3]，以及文化和语言偏见等问题。

对于人工智能的内容风险，我国政府高度重视并适时出台了一系列办法和要求。2022年11月25日，国家互联网信息办公室等三部门发布了《互联网信息服务深度合成管理规定》[4]，要

[作者简介]李栋，男，高级工程师，研究方向为信息资源建设与管理，Email:lidong@cass.org.cn。

求深度合成服务信息必须弘扬社会主义核心价值观，维护国家安全和社会公共利益，保护公民、法人和其他组织的合法权益。国家互联网信息办公室等七部门于 2023 年 7 月发布的《生成式人工智能服务管理暂行办法》[5]，对生成式人工智能服务提出明确的价值观和安全要求。这些制度和要求展现出国家在重视人工智能技术发展的同时，也在强化对其安全监管，尤其是对价值观和意识形态的监管。本文通过分析生成式人工智能交互内容在舆论传播、文化认同、意识形态等方面的潜在风险，挖掘其背后语料库作为大语言模型训练基础对生成式人工智能的重要影响，对其目前发展存在的瓶颈与困难进行分析，提出关于人工智能语料库建设的建议。

1　生成式人工智能交互内容风险分析

国外生成式人工智能产品利用语料训练形成的类似专家顾问式的交流及其广泛的社会影响，在一定程度上对社会大众理解判断社会现象、社会关系和社会规范进行了重构，这种影响对我国争取有利国际舆论环境、获得更为广泛的国际话语权以及防范意识形态领域风险等带来挑战，其风险主要体现在以下四个方面。

1.1　公众舆论误导风险

生成式人工智能产品在交互式对话中形成的看法、意见和评判等会潜移默化地引导公众舆论，在更大范围对用户认知产生影响，尤其在“马太效应”的作用下，会进一步扩大舆论传播优势。在生成式人工智能技术支持下，操控者可以通过大量人工智能虚假账号参与某个网络议题表态，构建虚假参与环境，夸大主题影响，吸引大量真实用户参与，操弄人群的热情和认识，鼓动其在现实社会付诸行动[6]。如果生成交互信息包含虚假、有害内容，将在舆论上给社会公众带来误导，引发公众思想的混乱，甚至给社会带来不安定因素。2023 年 5 月，甘肃平凉公安机关破获的利用 ChatGPT 炮制虚假信息并在网上大规模散布的案件就是此类风险的典型案例[7]。

1.2　文化认同风险

目前，ChatGPT 等国外模型训练语料主要取自英文语料库，来源主要包括维基百科、新闻报道、社交媒体、电子书籍、论坛帖子等，中文比重不足千分之一，英文语料占比超过 92.6%[8]。由于中西文化的源头、演进路径不同，导致国外生成式人工智能产品可能对中华文化的包容性和多样性的认识不充分，生成交互信息的文化内涵和精神往往带有局限性和片面性；同时，大模型对不同国家地区用户输入内容的语言习惯、表达方式在理解和处理方面也可能存在较大的差异，从而进一步加大误解风险。因此，如果国内用户使用中文与 ChatGPT 交流，潜移默化地受到其语料库内容文化影响，可能对自己所属文化或民族的价值、地位、意义等产生怀疑、否定甚至排斥的心理。

1.3　增强国际话语权的挑战

考虑到 ChatGPT 等国外大模型对于基础语料筛选，不可避免体现西方价值观，忽视来自其他数据内容的不同立场，使得西方价值观强化其在国际话语权上的既有优势。部分青年群体在使用生成式人工智能模型过程中，易受 AI 技术新鲜感吸引而盲目听从、随声附和，难以识别隐含的

偏见谬误，将可能对我国社会认知、社会伦理和文化理念等带来不利影响，进而冲击我国主流价值观。

1.4　意识形态风险

生成式人工智能独特的对话式话语生成能力使其比脸书、推特以及微信等社交工具拥有更为强大的社交影响力。在少数国家技术垄断的环境下，ChatGPT 极有可能成为意识形态战争的工具。如果蕴含西方意识形态的内容通过隐性渠道渗透国内舆论领域，导致“普世价值”、历史虚无主义等错误观点甚嚣尘上，将会对我国意识形态完全造成威胁。

2　我国人工智能语料库建设的瓶颈和限制

西方国家在长期发展中形成了较为成熟的英文语料库体系，并通过在人工智能产品技术上的领先优势不断渗透到中文语料领域，对国内中文语料库形成挑战。为争夺语料资源阵地和主导权，在人工智能时代来临之际占据有利位置，必须站在国家战略的高度，加大力度建设我国自己的人工智能语料库。目前，我国人工智能语料库建设虽然取得一定发展成果，但总体水平还处于相对落后的状态，要全面推动语料库建设向深度、广度拓展，必须破解其发展中所面临的诸多瓶颈。

2.1　人工智能语料库的统一规划问题

传统意义上的语料库是由大量在真实情况下使用的语言信息集成的专供研究使用的资料库，主要应用于语文教学、语言研究、语言文字规范标准制定、辞书编纂、语言信息处理等方面[9]。西方国家英文语料库建设相对比较成熟，如英国国家语料库（British National Corpus, 简称 BNC）[10]、美国当代英语语料库（Corpus of Contemporary American English，简称 COCA）[11]以及柯林斯英语语料库（the Bank of English，简称 BOE）、美国国家语料库（American National Corpus，简称 ANC）等，这些语料库具有容量大、权威性高、类型丰富、时效性强等特点，有的提供便捷的检索和比较功能，COCA 等还可以免费在线使用，这些特点都为我国语料库建设提供了一定的借鉴。人工智能语料库与传统语料库不同，更倾向互联网语境下的自然语言处理。目前，国内在中文领域建设了一批语言类通用语料库及在线平台，如北京大学中国语言学研究中心的 CCL 语料库[12]、北京语言大学语言智能研究所的 BCC 汉语语料库[13]、教育部语言文字应用研究所的国家语委现代汉语语料库[14]、中国语言资源联盟[15]等。此外，在编程、医疗、传媒等垂直领域，侧重建设了一批开源中文通用和专业数据集。但是，国内的语料库在数量、质量方面与国外相比还存在一定差距，国内人工智能语料库建设整体上还较为松散，深度的、系统的统筹、整合和突破还不充分，相关指导意见和规范也尚未明确提出。

2.2　人工智能语料库的规范标准问题

传统的语料库建设规范可以分为编码规范和内容标注规范。国内语料库编码规范常使用行业惯例，较少使用国际标准，在一定程度上不利于语言资源的共享。在文本语料内容标注领域，国内主要出台了推荐性标准《信息处理用现代汉语分词规范》[16]、《信息处理用现代汉语词类标记规范》[17]等，但在实际工作中应用程度还不够高，可能带来语料库的异构问题，使得各库之间数据交换难以实现。在行业专用语料库标准方面，一些行业组织出台了相关成果，如中国翻译协

会于 2009 年发布的中国语言服务行业规范《语料库通用规范》[18]，描述并规定了语料库的建设与加工、管理与维护、交易与共享等方面的基本框架，但通用的元数据标注、内容标注和语料库评价的具体规范仍需进一步探讨和细化[19]。2023 年中国翻译协会发布了《中国特色话语翻译高端语料库建设》系列标准[20]，这些标准侧重于在语言资源、语言教学与培训、语言翻译、语言技术工具开发以及语言相关咨询业务等领域，但是对人工智能领域语料的范围、内容、标识等要求的针对性还不足，难以支持人工智能语料库的建设工作。

2.3 人工智能语料库的商业利益和知识产权限制

语料库建设需要汇聚各行各业数据，一些商业数据平台虽然收录范围广、社会认可度高，但出于商业利益考虑，在长期发展过程中已经形成自己的数据壁垒，数据开放、共享共用难度较大，难以作为语料基础平台。国内不少商业大模型团队利用私有数据库建设数据集，例如百度的内容生态数据、腾讯的公众号数据、知乎的问答数据、阿里的电商和物流数据等，这些花费不菲的商业数据集难以作为开源资源免费提供给其他组织机构共享。此外，语料库内容还受制于知识产权保护，一些是由于语料库设计不合理导致，比如因收录大量全文而引起的版权问题，限制了语料库的对外开放[21]，另一些问题则与版权保护制度有关[22]。现实中语料内容的知识产权保护会对语料库建设和共享造成一定限制，语料库内容越庞大，因版权问题产生的经济、法律代价就越高。

2.4 人工智能语料库的文化传承和舆论传播问题

语料库作为重要的历史资源、文化资源、语言现实生活资源，在建设过程中需要面对的是承载五千多年文明史的语言文字宝库，而不是仅在学术、技术或商业上的意义。目前，语料库建设中对自身承载的中华民族的精神内核、文化基因尚缺乏充分认识，需要从坚定文化自信，建设文化强国的角度，确保语料内容符合我国主流思想舆论要求，守住意识形态安全底线。语料建设天然带有传播特性，在人工智能和互联网时代，时空、地域的界限已经逐渐淡化，网络世界的舆论权争夺、意识形态斗争日益严峻，强化传播能力建设，遵循“七个着力”重大要求[23]，成为语料库舆论传播作用必须要考虑的重要因素。

3 建设人工智能语料库的建议

对于生成式人工智能产品而言，训练语料库直接影响交互内容的立场和倾向，只有从国家层面确立语料库正确的政治方向、价值取向和学术导向，突出语料库的公益开放和国际化特色，才能形成具有中国特色的人工智能语境，才能为人工智能时代的到来奠定数据基础。下面从四个方面提出关于新时代中国特色中文人工智能语料库建设的建议。

3.1 完善语料库政策制度设计

建设国家级人工智能语料库，充分发挥国家平台的指引性、权威性、基准性作用。在建设过程中，由国家相关机构牵头，制订相关指导意见，建立数据标准体系、审核评估体系和基础校核机制。在国际标准和行业认同的基础上，建立统一的数据标准体系，形成统一的语料库编码和标注规范，有机汇集国内已有语料库，实现多源数据集成，解决数据孤岛问题，促进开放共享。严

格审核机制和相关技术手段，强化内容监管，过滤、屏蔽违法不良信息，防止数据污染。发挥国家平台基础库的作用，作为国内外其他中文语料库内容比对基础，实现对语料内容观点、立场的审核校准，确保内容符合主流价值观、国家安全要求和社会公共利益要求。

3.2　坚持语料内容正确的政治方向

语料库是有意识形态属性的，只有将主动权掌握在自己手中，才能避免国外人工智能大模型对国内的渗透，避免蕴含西方价值观的所谓“人工智能生成结果”对国内舆论场和学术圈蒙蔽、误导。入库语料要严格审核把关，牢牢把住政治标准这一硬约束、硬要求，可以通过新时代党建话语体系来指导语料库内容建设，尤其是有关哲学社会科学领域的相关论述、论断，用中国共产党的党建思想理论和知识体系的话语呈现或表达中国共产党人的思想观念和价值追求[24]。语料内容要全面展示习近平新时代中国特色社会主义思想的最新研究成果，时政类数据可以重点吸纳中共中央宣传部打造的“学习强国”平台、人民网的“人民数据”及其他主流媒体的时政数据平台或数据集。坚持语料库的历史属性，具有五千多年悠久历史的中华文明，起源与发展具有本土性。语料内容要反映党的十八大以来国家建设发展的重要成果，传承弘扬中华优秀传统文化，要高度关注古籍、年鉴、方志等历史文献及其相关论文内容审核，体现唯物主义历史观的引领作用。

3.3　坚持语料库的公益开放特色

建设国家级人工智能语料库，要坚持面向世界、公益免费、开放获取的特点。充分借助开放科学这一重要的发展模式和全球共识，采用知识共享许可制度等数据资产使用和发布的风险防控措施[25]，保障许可方和被许可方的权益，在一定程度上避免因版权问题而导致的法律风险，让使用者和开发者自由使用和分发数据。对于现有的一些语料库，通过灵活运用政策，形成有效知识产权管理机制，大幅减少著作权、商标、专利侵权以及侵犯商业秘密等问题发生。建设过程中，以国家级公益性学术平台为基础，做好各类开源数据和开放获取资源汇集、整合，特别要重视与公益性学术平台的整合，如为贯彻落实习近平总书记在哲学社会科学工作座谈会上重要讲话精神，由中国社会科学院牵头建设的国家哲学社会科学文献中心，目前已收集中文期刊数据1383余万条，在国内外具有广泛影响[26]。

3.4　建设面向世界的语料库

建设国家级人工智能语料库是一项长期性任务，应采取分阶段、分步骤的方式，前期以简体和繁体中文为主，后期逐渐推动中文语料库向英文等其他语种语料库拓展。突出语料库的国际传播能力建设，在世界范围的各类语料数据平台、论坛及社区上推荐、推广我国语料库，为全球各类人工智能模型训练提供好用、方便的训练数据，逐步使之成为全球中文领域语料训练的基础数据。

在人工智能时代，建设具有新时代中国特色的人工智能专业语料库，是中国哲学社会科学工作者的光荣使命，通过赋予人工智能系统具有权威性、公益性和开放性的语料库，才能确保我国主流意识形态在人工智能时代不失语、不失声，才能在人工智能时代来临之际占据有利位置。

【参考文献】

［1］张庆国. 生成式人工智能内容安全风险分析与安全机制探讨［J］. 人工智能，2024（2）：79-86.

［2］郭小东. 生成式人工智能的风险及其包容性法律治理［J］. 北京理工大学学报（社会科学版），2023（6）：93-105，117.

［3］蓝江. 生成式人工智能与人文社会科学的历史使命：从ChatGPT智能革命谈起［J］. 思想理论教育，2023（4）：12-18.

［4］中国网信网. 互联网信息服务深度合成管理规定［EB/OL］.［2022-12-11］. http：//www. cac. gov. cn/2022-12/11/c_1672221949354811. htm.

［5］中国政府网. 生成式人工智能服务管理暂行办法［EB/OL］.［2023-07-10］. https：//www. gov. cn/zhengce/zhengceku/202307/content_6891752. htm.

［6］向征. “黑镜”中的对垒：生成式人工智能背景下网络意识形态风险与防范［J］，社会科学战线，2024（4）：18-24.

［7］澎湃新闻. 甘肃警方：男子用ChatGPT编造虚假信息被采取刑事强制措施［EB/OL］.［2023-05-07］. https：//www. thepaper. cn/newsDetail_forward_22992435.

［8］张田勘. 破解大模型中文语料不足问题，并非毫无办法［EB/OL］.［2024-03-11］. https：//www. bjnews. com. cn/detail/1710142219169246. html.

［9］张伯江，张永伟. 国家语料库是重大文化资源［EB/OL］.［2023-11-08］. https：//www. cssn. cn/skgz/bwyc/202311/t20231108_5695444. shtml.

［10］英国国家语料库BNC［EB/OL］.［2024-07-12］. https：// http：//www. natcorp. ox. ac. uk/.

［11］当代美国的语料库COCA［EB/OL］.［2024-07-12］. https：//www. english-corpora. org/coca/.

［12］北京大学中国语言学研究中心. 北京大学现代汉语语料库［EL/OB］.［2024-05-18］. http：//ccl. pku. edu. cn：8080/ccl_corpus/index. jsp.

［13］北京语言大学语言智能研究所. BCC语料库［EB/OL］.［2024-05-18］. https：//bcc. blcu. edu. cn/.

［14］教育部语言文字应用研究所. 中国语言文字网［EB/OL］.［2024-05-18］. http：//www. china-language. edu. cn/.

［15］中国语言资源联盟. 中国语言资源联盟［EB/OL］.［2024-05-18］. http：//www. chineseldc. org/.

［16］北京航空航天大学等. 信息处理用现代汉语分词规范：GB/T 13715-1992［S］. 北京：中国标准出版社，1992：10.

［17］教育部语言文字应用研究所. 信息处理用现代汉语词类标记规范：GB/T 20532-2006［S］. 北京：中国标准出版社，2006：9.

［18］中国翻译协会. 语料库通用技术规范：ZYF 001-2018［S］. 北京：中国翻译协会，2019：1.

［19］钱小飞. 语言数据资源建设中的关键问题及对策［J］. 语料库语言学，2021，8（2）：94-105.

［20］中国翻译协会. 中国特色话语翻译 高端语料库建设 第一部分：基础要求：T/TAZ7. 1-2021［S］. 北京：中国翻译协会，2022：5.

［21］肖忠华. 肖忠华语料库语言学答客问［J］. 语料库语言学，2015，2（2）：1-14，115.

［22］程亚丽，王海萍. 我国英语语料库的版权保护研究［J］. 编辑之友，2012（11）：87-90.

［23］新华网. 习近平对宣传思想文化工作作出重要指示强调［EB/OL］.［2023-10-08］. http：//www. news. cn/politics/2023-10/08/c_1129904890. htm.

［24］赵绪生，孙进宝. 论新时代党建话语体系建设［J］. 中共中央党校（国家行政学院）学报，2020，24（5）：38-46.

［25］雷聪仪，顾立平，聂华，等. 开放获取的许可授权协议管理［J］. 中华医学图书情报杂志，2021，30（3）：17-23.

［26］赵以安，赵语嫣，李菲菲，等. 推动学术发展　勇担文化使命　不断加快国家哲学社会科学文献中心建设——国家哲学社会科学文献中心关注度报告发布会综述［J］. 文献与数据学报，2024，6（2）：3-13.

Thoughts on Building a Chinese AI Corpus with Chinese Characteristics for the New Era

Li Dong

(Chinese Academy of Social Sciences Library, Beijing 100732, China)

Abstract:［**Purpose/Significance**］To seize the initiative, guidance, and discourse power in international public opinion, great importance must be attached to the construction of Chinese AI corpus.［**Method/Process**］By analyzing the potential risks of AIGC interactive content in public opinion dissemination, cultural identity, ideology, etc., this study investigates the bottlenecks and difficulties in the current development of AI corpus. Based on this, it proposes a development path for Chinese AI corpus construction.［**Result/Conclusion**］Advancing the construction of Chinese AI corpus with national strategic coordination for the new era, swiftly overcoming restrictive bottlenecks, and addressing the development from four perspectives: refining corpus policy design, adhering to the correct political orientation for corpus content, upholding the public and open nature of the corpus, and building a corpus oriented towards the world. This approach will lay the data foundation for the advent of the AI era.

Keywords: Artificial Intelligence (AI); Corpus; Public benefit and open

（本文责编：孔青青）